DOCUMENT RESUME

ED 429 309 CS 216 672

AUTHOR	Valencia, Sheila W.; Wixson, Karen K.
TITLE	Policy-Oriented Research on Literacy Standards and Assessment.
INSTITUTION	Center for the Improvement of Early Reading Achievement, Ann Arbor.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO	CIER-3-004
PUB DATE	1999-04-01
NOTE	42p.
CONTRACT	R305R70004
AVAILABLE FROM	CIERA/University of Michigan, 610 E. University Ave., 1600 SEB, Ann Arbor, MI 48109-1259 (printed copy: $6.50 including handling and shipping); full report may be downloaded from Web site: www.ciera.org
PUB TYPE	Information Analyses (070)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Educational Change; *Educational Policy; Elementary Secondary Education; *Evaluation Methods; Literature Reviews; *Reading Achievement; Reading Instruction; *Reading Research; Research Needs; Writing Instruction; *Writing Research

ABSTRACT

Mandates about the kind and amount of literacy instruction in classrooms and levels of reading required for successful promotion from grade to grade are currently dominating state and federal legislation. This paper gathers scholarship from three disparate sources to produce the first integrated review of policy-related research on literacy education: (1) policy analyses that examine policies about literacy in the framework of systemic reform; (2) measurement and evaluation studies, conducted by psychometricians, that examine the assessments mandated by policies; and (3) studies by literacy researchers that attend to policies and literacy-specific content. Each of these literatures has a unique set of questions, frameworks, methodologies, and audiences. Together, they provide a comprehensive perspective on policies and literacy education. The paper concludes that the best way to influence policy, instruction, and children's learning is for policy, measurement, and literacy researchers to collaborate in conducting and reporting research. Each group needs to learn more about the others' work in order to affect real change in literacy practice. Contains 121 references. (RS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 429 309

CIERA REPORT

Policy-Oriented Research on Literacy Standards and Assessment

Sheila W. Valencia, University of Washington
Karen K. Wixson, University of Michigan

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

CIERA REPORT #3-004
April 1, 1999

CIERA
Center for the Improvement of Early Reading Achievement

CS 216 672

BEST COPY AVAILABLE

Policy-Oriented Research on Literacy Standards and Assessment

CIERA REPORT #3-004

Sheila W. Valencia, University of Washington
Karen K. Wixson, University of Michigan

CIERA Inquiry 3: Policy and Profession
How do the policies and frameworks of states guide local curriculum, instruction, and assessment at early grade levels? What is the impact of standards and assessment on literacy practice and student learning?

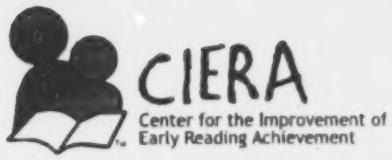
Mandates about the kind and amount of literacy instruction in classrooms and levels of reading required for successful promotion from grade to grade are currently dominating state and federal legislation. Valencia and Wixson have gathered scholarship from three disparate sources to produce the first integrated review of policy-related research on literacy education: (a) policy analyses that examine policies about literacy in the framework of systemic reform; (b) measurement and evaluation studies, conducted by psychometricians, that examine the assessments mandated by policies; and (c) studies by literacy researchers that attend to policies and literacy-specific content. Each of these literatures has a unique set of questions, frameworks, methodologies, and audiences. Together, they provide a comprehensive perspective on policies and literacy education.

The best way to influence policy, instruction, and children's learning, say Valencia and Wixson, is for policy, measurement, and literacy researchers to collaborate in conducting and reporting research. Each group needs to learn more about the others' work in order to affect real change in in the lives of children and teachers.

CIERA

Center for the Improvement of Early Reading Achievement
University of Michigan – Ann Arbor

April 1, 1999



University of Michigan School of Education

610 E University Ave., Rm 1600 SEB
Ann Arbor, MI 48109-1259

734.647.6940 voice
734.763.1229 fax
ciera@umich.edu

www.ciera.org

Chapter forthcoming in Kamil, M. L., Pearson, P. D., & Mosenthal, P. (Eds.), *Handbook of Reading Research* (3rd. ed.). Mahwah, NJ: Erlbaum.

©1999 Center for the Improvement of Early Reading Achievement.

This research was supported under the Educational Research and Development Centers Program, PR/Award Number R305R70004, as administered by the Office of Educational Research and Improvement, U.S. Department of Education. However, the comments do not necessarily represent the positions or policies of the National Institute of Student Achievement, Curriculum, and Assessment or the National Institute on Early Childhood Development, or the U.S. Department of Education, and you should not assume endorsement by the Federal Government.

Policy-Oriented Research on Literacy Standards and Assessment

Sheila W. Valencia
University of Washington

Karen K. Wixson
University of Michigan

In the past two decades, instruments of policy have reached into every facet of our educational lives. The "tools" of policy include everything from new content standards and instructional frameworks to teacher certification requirements, systems of assessment, Title I allocations and requirements, and textbook adoption guidelines. This report focuses on a discussion of policy-oriented research on literacy content standards and assessment. Understanding how policy instruments such as standards and assessments have attained such importance in subject matter learning provides important background underlying much of the research.

Historically, state policymakers have delegated their authority over public education to local school districts, particularly in matters of curriculum and instruction. Districts, in turn, have entrusted the curriculum to teachers or indirectly to textbook publishers, and done little to provide instructional guidance (Massell, Kirst, & Hoppe, 1997). Since the publication of the now-famous report *A Nation at Risk* (National Commission on Excellence in Education, 1983), however, states and districts have made unprecedented forays into curriculum and instruction (Massell, Kirst, & Hoppe, 1997). This modern reform movement has been characterized by efforts to create new "policy instruments" to elicit, encourage, or demand changes in teaching and learning, and reduce the tangles of regulation, bureaucracy, proliferating policy, and incoherent governance that would impede reform (Smith & O'Day, 1991). Included among the new policy instruments are the standards and assessments that are the subject of the research we examine here.

As we considered what literature to include in this review, we were aware that current reform efforts have contributed to an increased interest in policy research. For example, the Office of Educational Research and Improvement (OERI) funded the Consortium for Policy Research in Education in 1985 and the Center for the Study of Teaching and Policy in 1998, and the American Educational Research Association established Division L on politics and policymaking in 1997. In addition to the growing number of policy researchers, researchers in the areas of measurement and evaluation have

become interested in policy because many reform initiatives have focused on assessment as a primary vehicle for improving student achievement. Similarly, the many reform efforts aimed at improving the literacy levels of young Americans have led literacy researchers to become more interested in research on policy-related issues. The realization that policy-oriented research is being conducted from a variety of perspectives led us to approach this review on two levels. On one level, we characterize the nature of policy-oriented research on literacy standards and assessments. On another level, we review what research tells us about the impact of literacy standards and assessment on practice and student learning.

To characterize the nature of literacy policy research, we examine the literature on standards and assessment in relation to the policy, measurement, and literacy contexts from which it arises. As a result, we review three fairly distinct sets of literature. Policy researchers generally set out to address policy issues head-on, and are less concerned with subject-matter specifics. Measurement researchers are also more concerned with general findings than with subject matter specifics, but they tend to focus on the qualities and influence of assessment policy tools rather than on policy questions per se. Literacy researchers are rarely driven by policy questions or issues; they are primarily interested in subject matter teaching and learning. Differences in perspective suggest different research questions, conceptual frameworks, methodologies, perspectives on literacy, and audiences for publications, which result in differences in what is learned about literacy standards and assessment.

In order to present these different perspectives and findings clearly, we have organized this report into four sections. The first three sections present our review of the policy-oriented literature related to literacy standards and assessment in terms of the three research perspectives—policy, measurement, and literacy. Each of these sections has two parts: a brief discussion of background, situating the research in its larger context, and a review of the literacy policy research within each perspective. The report's fourth section focuses on what we have learned from these bodies of research with regard to the nature of the research and the results of policy-oriented research on literacy standards assessment.

The Policy Perspective

Following the publication of *A Nation at Risk* (National Commission on Excellence in Education, 1983), two "waves" of school reform emerged (Lusi, 1997). The first wave involved state efforts to accomplish three goals: (a) to raise coursework standards for high school graduation, (b) to implement and/or expand assessment programs, and (c) to raise standards for prospective teachers (Goertz, Floden, & O'Day, 1995). The second wave of reform took the form of school restructuring, and combined three complementary elements: (a) a call for higher and common expectations for ALL students, (b) an emphasis on new and more challenging teaching practices, and (c) dramatic changes in the organization and management of public schools (Elmore, 1990).

These initial waves of reform did little to change the content of instruction, especially with their focus on basic skills. Nor did they result in the desired changes in teaching, learning, and student achievement (Cuban, 1990; Firestone, Fuhrman, & Kirst, 1989). Fragmented and contradictory policies diverted teachers' attention, provided little or no support for the type of professional learning they needed, and made it difficult to sustain the very promising reforms taking shape in individual schools or clusters of schools (Cohen & Spillane, 1992; Goertz et al., 1995).

Growing concerns about the educational preparation of the nation's youth prompted President Bush and the nation's governors to call an education summit in Charlottesville, Virginia, in September, 1989. This summit resulted in the establishment of six broad educational goals to be reached by the year 2000 (National Education Goals Panel, 1991). In pursuit of the National Education Goals, the bipartisan National Council on Education Standards and Testing (NCEST) issued a report in January, 1992, recommending national content standards and a national system of assessments based on the new standards. Precedent for and guidance in developing national standards was to be found in the work of the National Council of Teachers of Mathematics (NCTM), published as Curriculum and Evaluation Standards for School Mathematics in 1989. The logic was that once broad agreement had been achieved on what is to be taught and learned, then everything else in the system (e.g., tests, professional development, textbooks, software, etc.) could be redirected toward reaching those standards. This view has come to be known as "systemic reform."

Systemic reform advocates changing teaching as the most direct way to change students' learning (Cohen, 1995). and it is posited as a way to provide top-down support for bottom-up instructional improvement in classrooms, schools, and districts. The key question for reformers has been how to get there—how to foster (or mandate) changes in learning and teaching. Many systemic reformers have seen government as their chief vehicle. However, efforts by groups such as the Coalition for Essential Schools, the Accelerated Schools Network, and the New Standards Project—groups which have substantial resources but operate largely outside the framework of governmental policy—indicate that state and federal policies are not the only ways to pursue improved instruction. Systemic reformers have tended to focus on creating new policy instruments such as content standards or curricular frameworks, assessments that are aligned with new content standards, and changes in both preservice and inservice teacher education (Cohen, 1995).

According to McLaughlin (1987), policy research into the late 1980s generated a number of important lessons for policy, practice, and analysis by acknowledging the role of contextual factors such as local priorities, individual beliefs and motivation, and the balance between support and pressure to change. Furthermore, McLaughlin saw these lessons as framing the conceptual and instrumental challenge for the next generation of policy analysts—to describe a model of implementation highlighting individuals rather than institutions and viewing implementation issues in terms of individual actors' incentives, beliefs, and capacities. Darling-Hammond (1990) added that top-down policies could "constrain but not construct" change. She focused on policy enactment, arguing a) that local leadership and motivations for change are critical to policy success; b) that local agencies must adapt poli-

cies rather than adopting them because local ideas and circumstances always vary; and c) "that teachers' and administrators' opportunities for continual learning, experimentation, and decision making during implementation determine whether policies will come alive in schools or fade away when the money or enforcement pressures end" (p. 235).

At a deeper level, Darling-Hammond (1990) argued that we knew little about the meaning of specific policies for educational life within classrooms. She indicated that advances in policy analysis during the 1980s made it possible to ask a number of new questions. For example, what differences do such advances actually make to teachers' and students' work together? How do teachers understand and interpret the intentions of new policies in the context of their knowledge, beliefs, and teaching circumstances? How and under what circumstances do policies that are intended to change teaching actually do so? These observations were presented in Darling-Hammond's (1990) introduction to a special issue of *Educational Evaluation and Policy Analysis* (EEPA) focused on case studies of California reform in K-12 mathematics education. These case studies were seen as leading the way toward a new generation of policy analysis that recognized "the importance of understanding the transformation of policy into teacher actions from the vantage point of the teachers, themselves, as well as from that of the policy system" (p. 175).

Our review of policy research in the 1990s revealed relatively few studies that clearly addressed literacy standards and assessment. The few that did could be broken into two types. First, there were large-scale investigations of state reform efforts. These investigations began to link macro- and microlevels of analysis using classroom artifacts such as lesson assignments and interviews with teachers and administrators to get at the classroom perspective. Second, there were investigations, often case studies, that explored more deeply the impact of policy instruments on teachers, schools, and districts. By limiting our review to policy research related to literacy standards and assessments, we saw increased efforts to examine policy initiatives with what Darling-Hammond called a "pedagogical eye," but little attention given to the role that different subject areas might play in implementation (although there is some indication that this too may be changing; see Ball & Cohen, 1995).

Research

Large-scale investigations of state reform efforts.

A study by Goertz, Floden, and O'Day (1995) provides an example of how large-scale policy research projects in the 1990s have begun to link macro- and microlevels of analysis. The stated purposes for this study included expanding our knowledge of state approaches to education reform; examining district, school, and teacher responses to state reform policies in a small number of reforming schools and districts; and studying the capacity of the educational system to support education reform. The findings are based on case studies of 12 reforming schools located in six reforming school districts in three states that have taken somewhat different approaches to systemic reform—California, Michigan, and Vermont. The researchers interviewed educators, administrators and policymakers at the school, district, and state

levels. They also surveyed and interviewed 60 teachers in each school. Because it was too early in the reform movement to assess the impact of any particular state, district, and/or school strategy, this study was intended more as a description of what was happening than "what works."

It is noteworthy that not until halfway through the seven-page Executive Summary do Goertz et al. mention that this research focused on reform in mathematics and language arts. Since we are concerned here with literacy related policy research, we summarize only the section of the report that focuses specifically on language arts policy. From a policy researchers' perspective, however, we should remember that this report is not about mathematics or language arts reform. It is about systemic reform, and the attention given to mathematics and language arts merely provides some specificity to the findings.

The portion of the report that deals directly with language arts examines the degree to which teachers' reports on their instruction were consistent with explicit or implicit curriculum recommendations set forth in curriculum frameworks and state assessments. For example, some of the aims of California's "meaning-centered" English-language arts reform were reflected in survey data on reading instruction. Elementary teachers reported that, during reading instruction, their students spent three and one-half hours per week on comprehension strategies and responding to why they read. The least amount of time was spent on word recognition skills (30 minutes) and phonics (19 minutes). The California Department of Education's English Language Arts Framework (1987) also emphasized a literature-based curriculum that "engaged students with the vitality of ideas and values greater than those of the marketplace or video arcade" (p. 7). Elementary teachers indicated that 80 percent of instructional time was spent using literature trade books, and the remaining time was distributed among reading or subject basals, workbooks or worksheets, or something else.

As in California, Michigan teachers emphasized content matching the "meaning-centered" view of Michigan's Essential Goals and Objectives in Reading (1986). For example, both elementary and secondary teachers in Michigan reported spending over three hours per week on comprehension strategies and having students respond to what they read, and slightly more than one-half hour per week on basic skills, such as phonics and word recognition. Both California and Michigan teachers reported spending roughly the same amount of time per week on reading instruction. However, California teachers spent over four and one-half hours per week with students engaged in small group reading activities, such as working in pairs or teams and small group discussions. In contrast, Michigan teachers spent less than two and one-half hours per week engaging in these kinds of activities. These differences in instructional practices reported by Michigan and California teachers are consistent with the emphasis given to dissimilar aspects of the language arts reform policies in the two states.

Goertz et al. concluded that there is evidence of general patterns that incorporate new directions in both state and national reforms but also retain attention to more traditional topic areas. Teachers believed that they had been influenced by state policy instruments, such as assessments and curricular frameworks, but that these state influences were by no means the only or even the most important influences on practice. Teachers reported that their

own knowledge and beliefs about the subject matter and about their students generally had a larger influence on their teaching than state policies.

Other examples of policy research that reflects initial efforts to link macro- and microlevels of analysis include the work of McDonnell (McDonnell, 1997; McDonnell & Choisser, 1997), Smith and colleagues (Smith, Noble, Heinecke, Seck, Parish, Cabay, Junker, Haag, Taylor, Safran, Penley, & Bradshaw, 1997), and the Kentucky Institute for Education Research (Lindle, Petrosko, & Pankratz, 1997), all of whom studied state reforms that included a literacy component. For example, McDonnell and Choisser examined the extent to which policymakers' expectations about the curricular effects of testing in Kentucky and North Carolina proved valid in local schools and classrooms. Their analysis was based on telephone and on-site interviews with teachers and administrators, and examinations of assignments and daily logs gathered from 23 teachers in each state. They concluded that transforming instruction through assessment was not a self-implementing reform because the tests alone lacked sufficient guidance for how teachers ought to change.

Smith et al. took a different methodological approach, conducting a four-year, multimethod approach to study the effects of the now suspended Arizona Student Assessment Program (ASAP). They observed in classrooms and interviewed teachers and principals in four schools, and they used a survey approach to collect data from educators across the state. The results, which were generalized across subject areas, indicated that though educators knew about ASAP, their responses to it varied depending on how they understood it and how it fit with their underlying beliefs and the local conditions (material and knowledge resources, existing beliefs and ideologies about teaching, culture of accountability and authority). Smith et al. also concluded that the dual focus on accountability and instructional improvement, combined with insufficient attention to capacity building, resulted in marginal effects of the ASAP reform agenda.

Investigations into the impact on teachers, schools, and districts.

In contrast to the large scale investigations represented by the Goertz et al., McDonnell, and Smith et al. studies, some policy researchers have conducted more in-depth studies of the impact of policy instruments on teachers, schools, and districts. For example, Standerford (1997) studied two small districts in Michigan from 1988 to 1991. Both districts formed reading curriculum committees in an effort to interpret the state reading policy and design an official district response. To understand what happened in these districts, Standerford observed both the curriculum committees and the classroom practices of the teachers on these committees. Her results indicated that participation in the district effort was not integrated with either state policy or the classroom changes that individual teachers were making; the district rules, objectives, players, audiences, and time frames were not conducive to such integration. Districts' responses to state reading reform were influenced by their need to reduce uncertainty, use standard operating procedures to effect change, advertise change by producing documents and plans, and respond selectively to policies based on the incentives attached. In contrast, the teachers made changes based on their individual professional development activities, but were often unsure just how those changes fit with the state policy.

Standerford concluded that state and district policies had influenced the teachers' efforts by making them aware that changes were expected in reading instruction, but had not made clear for the teachers what those instructional changes should be. Nor did these policies offer much support for teachers' efforts to figure that out for themselves. As teachers learned more about the new ideas, they gradually changed the enacted curriculum in their classrooms. Yet those instructional changes were minimally represented in the written curriculum that they produced as members of the district committees, because their roles and objectives were defined differently at the district and classroom levels.

In another series of studies, Spillane (1998), Spillane (1996), Spillane and Jennings (1997), and Jennings (1995) examined the impact of the reading policy in Michigan on a racially and economically diverse urban district, a relatively affluent suburban district, and a small group of teachers within the suburban district. Spillane's (1996) study revealed that state and local policies do not always support similar notions about instruction. The suburban district used the revision of the state reading test as a lever to move in another direction the central administrators who preferred a basic skills curriculum. Within a short period of time, district administrators had developed a new curriculum guide for reading, adopted new curricular materials, revised their student assessment policies, and organized an extensive professional development program about reading that went beyond state policy. In contrast, the state's reading policy did not figure prominently in the reading program developed in the urban district. Curriculum guides supported traditional ideas about teaching reading by encouraging teachers to teach isolated bits of vocabulary, decoding skills, and comprehension skills. A new basal reading program was mandated, accompanied by a traditional workbook that provided students with drill in reading skills. Central administrators made no effort to revise district policy on student assessment despite significant revisions of the state's reading test.

When Spillane and Jennings (1997) looked more closely at nine second- and fifth-grade teachers in the suburban district, they found that the extent to which teachers' practices reflected the district's literacy initiative depended on how well the reforms were elaborated by the district. Their initial data analysis suggested significant uniformity in language arts practice among the nine classrooms and offered striking evidence that the district's proposals for language arts reform were finding their way into practice. For example, they found that all nine teachers were using literature-based reading programs and trade books, engaging in activities such as Writer's Workshop, and focusing on comprehension over skills-based instruction. However, early discussions of the observation data revealed other differences that weren't being captured by the analytical framework. This led to a revised analytical frame focused on classroom tasks and discourse patterns that helped track these "below the surface" differences in pedagogy.

Comparing results using the two analytical frameworks, Spillane and Jennings showed that it is relatively easy to arrive at very different conclusions about the extent to which reforms that call for more ambitious pedagogy have permeated practice. They argue that, if reforms are meant to help all students encounter language arts in a more demanding and authentic manner, then policy analysts cannot rely solely on such indicators as the materials and activities teachers use. Rather, they must sit in classrooms and figure

out what type of knowledge is supported by classroom tasks and discourse patterns. We would add that to be able to explore these issues effectively, one needs to understand a great deal about the subject-matter instruction that is the focus of the reform. Spillane and Jennings have amassed a great deal of knowledge about language arts and language arts instruction over their years of studying reform in this area, and we would argue that, without this knowledge, they may never have even seen the differences that led them to revise their analytical framework and uncover these important differences in classroom practice.

Summary and Implications

Collectively, these studies provide insight regarding both policy research related to literacy standards and assessments, and the impact of literacy standards and assessments on district and teacher practices. On the one hand, very few policy studies provide sufficient subject matter information to warrant inclusion in this review. On the other hand, several studies probe deeply into the details of the classroom discourse and tasks related to language arts instruction, revealing important differences in teachers' implementation of reform efforts. In terms of the impact of literacy standards and assessments, it is clear that policy tools such as conceptual frameworks, curriculum guides, and assessments can and do influence district and classroom practice. It is also evident that the relations between language arts policy and practice are complex and at least partly dependent on the knowledge, beliefs, goals, and experience of the administrators and teachers who work with these types of policy tools. These findings speak clearly to the need to understand thoroughly the context of policy implementation from both the system perspective and the day-to-day lives of teachers and students. They also suggest that, without some form of professional development, the effects of such policies are highly variable.

The Measurement Perspective

Assessment has been part of educational reform efforts for the past 40 years (Linn, 1998), initially serving as an indicator of reform or progress and more recently serving as a lever for reform. In the 1960s, testing increased substantially to meet the demands of evaluation and accountability for Title I. Then in the 1970s and 1980s, measurement researchers became intimately involved in policy-related issues during the minimal competency testing (MCT) movement when high stakes were attached to test performance. In Florida, for example, where MCT graduation requirements gained a great deal of attention, test results revealed gains for low-achieving students but differential passing rates for African American, Hispanic, and white students. In addition, the Federal District Court decision in the landmark Debra P. vs. Turlington (1981) case directed that students must be provided with ample opportunity to learn the material tested when high stakes, such as high school graduation, are in place. Events such as these quickly propelled

assessment and assessment researchers into the policy arena. Following this trend, a new movement, measurement-driven reform (Popham, Cruse, Rankin, Sandifer, & Williams, 1985), gained in popularity, emphasizing large-scale assessment as a "catalyst to improve instruction" (p.628). Measurement-driven reform expanded the role of assessment into the policy arena in two important ways: a) it focused attention on what students should learn (outcomes), and b) it made teaching toward the test a valued instructional strategy.

Many measurement researchers explored the effects of early high-stakes assessments on student performance, curriculum, and teachers' instructional practices. In general, studies indicated that high-stakes standardized basic skills tests led to: a) a narrowing of the curriculum, b) an overemphasis on basic skills and test-like instructional methods, c) a reduction in effective instructional time and an increase in time for test preparation, d) inflated test scores, and e) pressure on teachers to improve test scores (Herman & Golan, 1993; Nolen, Haladyna, & Haas, 1992; Resnick & Resnick, 1992; Shepard, 1991; Shepard & Dougherty, 1991, Smith, 1991; Smith, Edelsky, Draper, Rottenberg, & Cherland, 1990). These studies led educators and the public alike to question the effectiveness of educational reform efforts and of the assessments themselves (Linn, Grau, & Sanders, 1990). As a result of this line of research and renewed interest in the intended and unintended consequences of assessment (Messick, 1989), the "alternative" or "authentic" assessment movement was launched.

From past research it was clear that assessment could be a lever for reform—that what gets tested gets taught, and what doesn't get tested doesn't get taught. Therefore, it was reasoned that if better, more authentic assessments could be created to measure the "thinking curriculum" (Resnick & Resnick, 1992), then better teaching and learning would follow. Publicly acknowledged content standards in specific subject areas would guide the content of the new assessments, and high performance standards, rather than norms, would guide goals for student achievement (NCEST, 1992). Furthermore, it was argued that if teachers were more involved in the development, administration, and scoring of the assessments, there would be a greater chance that teaching would be enhanced. Performance assessment, portfolios, and projects (Resnick & Resnick, 1992) were advanced by both educators and measurement experts as assessment models that might foster effective teaching, learning, and measurement of worthwhile outcomes (Shepard, 1989; Simmons & Resnick, 1993; Wiggins, 1993). In many respects, the authentic assessment movement is an extension of the measurement-driven reform of the 1980s. Now, however, the people involved in assessment development, the form of assessment, and the criteria for content selection and student performance have changed. Furthermore, there is new interest in students' opportunities to learn.

The measurement community cautioned that the field would require new models for and methods of determining the technical merit of new assessments (Linn, Baker, & Dunbar, 1991; Moss, 1994), many of which were not yet in place. Furthermore, many argued that it was impossible to test the logical assertion that these new measures would yield more positive results until the assessments were in place for some time. Nevertheless, pressure for new, better assessments and for public accountability placed new assessments on a fast track. By 1997, 46 out of 50 states had some form of state-

wide assessment, and 36 of those included extended responses typical of performance assessments (Roeber, Bond, & Braskamp, 1997). Linn (1998) has suggested that policy-makers have placed enormous emphasis on assessment reform because it is relatively inexpensive and easy to mandate, can be implemented rapidly, and is easily reported by the press, when compared to the type professional development and restructuring/reculturing of schools required to affect deep, second-order educational change (Fullan & Miles, 1992). So in the 1990s, we have witnessed an enormous increase in new assessments being used as the levers for reform.

The research we review in this section falls into two general categories: on-demand forms of performance-assessments and classroom-based assessments such as portfolios. We use the term "on-demand performance assessments" to define uniform assessments administered under controlled conditions; they are usually given on a particular day or days under standard conditions across classrooms, schools, and districts. Most statewide assessments in reading and writing are on-demand assessments. Recent efforts have focused on improving the quality of the assessment tasks and expanding response modes while simultaneously trying to maintain high levels of reliability and validity. In this section of the review, we include research on on-demand performance assessments that require students to demonstrate higher-order cognitive processes and to provide some extended responses to comprehension questions or to write in response to a prompt. We do not include research on more traditional assessments comprised only of multiple-choice items. For the second category—classroom-based assessments—we include assessment evidence that is systematically collected as an ongoing part of the instructional program. In some cases, the evidence is scored and then reported for accountability purposes either at the state or school-district level. Because we are focusing on policy-related research, we do not include research on individual classroom assessment projects.

Research

On-demand performance assessment.

The first attempts at performance assessment in literacy can be traced back to the 1960s and the use of direct writing assessment instead of indirect measures such as multiple-choice tests (cf. Freedman, 1993). Direct writing assessment requires students to write in response to an assigned topic under timed conditions; papers are scored using a standard rubric. Many statewide assessments (Roeber, et al., 1997) and the National Assessment of Educational Progress still use this approach with considerable success. Measurement researchers have focused on issues of interrater reliability and generalizability with respect to scoring writing samples. Interrater reliability is generally high, although studies indicate it can vary from .3 to .91 (Dunbar, Koretz, & Hoover, 1992; Hieronymus, Hoover, Cantor, & Oberley, 1987; Welch, 1991). Measurement researchers seem to understand how to raise reliability to an acceptable level by implementing more extensive training of carefully selected scorers, more specific scoring guidelines, and the like (Mehrens, 1992; Miller & Legg, 1993). Issues of generalizability across modes of writing or even topics within modes are not as clear, however, and continue to present challenges for measurement experts (Dunbar, Koretz, & Hoover, 1992; Herman, 1991; Hieronymus, Hoover, Cantor, & Oberley,

1987). Language arts educators, however, are now raising questions regarding the authenticity of direct writing assessment and the validity of the results when students are required to write under these unnatural, on-demand conditions (e.g., constrained by time, topic, audience, and process) (Freedman, 1993; Lucas, 1988a, 1988b). Although these criticisms are appealing on the surface, Messick (1994), a noted measurement researcher, pointed out that concerns about both authenticity and directness need to be supported empirically rather than simply claimed. This is a good example of how differences in perspective shape the questions and the nature of the evidence sought.

Although direct writing assessment is still a mainstay of many assessment programs, more recent efforts at performance assessment in reading and writing go further, including longer and more complex reading selections from a variety of genres, higher level comprehension questions, extended written responses, and cross-text analyses. The few studies from a measurement perspective that are available on new statewide assessments (e.g., Maryland, Kentucky, Arizona) do not distinguish among reading, language arts, and mathematics in design or analyses, making it difficult for literacy educators to interpret the implications for curriculum, instruction, or research. For example, in two parallel studies, researchers at RAND (Koretz, Barron, Mitchell, & Stecher, 1996; Koretz, Mitchell, Barron, & Keith, 1996) used telephone and written surveys to examine the influence of the Maryland School Performance Assessment Program (MSPAP) and the Kentucky Instructional Results Information System (KIRIS)—both of which had assessments in several subject areas. By focusing only on responses of elementary teachers included in these reports, we can get some idea of language-arts-related results. Across both studies, teachers supported the new assessments, even in terms of encouraging reluctant teachers to change; however, they did not support the use of test results for accountability. On the positive side, teachers aligned curriculum with the assessments, especially spending more time on writing (a dominant response mode for both assessments), although they felt that more specific curriculum frameworks would be helpful. On the negative side, teachers reported spending considerable time in test preparation activities and a tendency to de-emphasize untested material. Data from both studies indicated that teachers' expectations rose for high-achieving students rather than low-achieving students and that teachers credited student gains to specific test practice and test familiarity rather than to true improvements in capabilities. These findings led the researchers to call for further research on issues related to the specificity of the frameworks, effects on equity, inflated test scores, and the validity of the measures.

One of the few studies of on-demand assessment to report specifically by subject area is based on data from the New Standards Project, a multistate effort designed to involve educators in the creation of state and district performance-based assessments in reading, writing, and mathematics (Resnick, Resnick, & DeStefano, 1993). This shared emphasis on new assessments and professional development involved teachers in the development, piloting and scoring of the assessments. Researchers found "moderate" interrater reliability for both the reading and writing sections of the test—too low to use for judging students or educational programs. More interestingly, reliability varied depending on the task being scored, the approach to calculating reliability (correlation or agreement), and the scoring method used (holistic or

a combination of analytic and holistic). Moreover, individual students' scores varied depending on the scoring method used. The researchers suggested that more intensive training, a more selective scoring team, clearer rubrics, and better exemplars might improve interrater agreement. These findings mirror the findings discussed earlier related to direct writing assessment.

Classroom-based assessment.

The classroom-based measurement research has largely been conducted on portfolios. Interest in portfolios and policy stemmed from an attempt to join the advantages of classroom-embedded assessment with the need for large-scale public accountability. From the beginning, many were leery of trying to accomplish both purposes with one instrument, but the advantages in terms of teacher development, instructional practice, and student engagement motivated educators to try (Aschbacher, 1994; Haney 1991; Mehrens, 1998; Valencia, 1991).

The most widely studied of the large-scale portfolio projects is the Vermont Portfolio Assessment Program, although more measurement researchers have studied the mathematics portfolio than the writing portfolio (Koretz, McCaffrey, Klein, Bell, & Stecher, 1993; Koretz, Stecher, & Deibert, 1992; Koretz, Stecher, Klein, & McCaffrey, 1994; Koretz, Stecher, Klein, McCaffrey, & Deibert, 1993). Because statewide assessment was new in Vermont, this project was conceptualized as a system that would take hold gradually—it would be decentralized and "a very long effort" (Mills & Brewer, 1988, as cited in Koretz et al., 1994). According to state officials, it was designed to support sound educational practice, encourage professional development of education, encourage local autonomy, and provide comparable information across schools. The writing assessment was designed to be administered in grades 4 and 8 (in 1994-95, the writing assessment was moved to grades 5 and 8) and is comprised of two main components in writing: a) a portfolio of student work which includes a set number and specified types of pieces of writing collected over the course of a year, and b) a "uniform test" of writing (i.e. a standard prompt to which all students respond using standard procedures). The portfolio contents and the Uniform Test are scored by a wide range of Vermont teachers other than the students' own, using an analytic scoring rubric.

Studies of reliability indicated interrater correlations ranging from .46 to .63 (45 percent agreement based on exact match) depending on how the scores were aggregated (e.g., within or across scoring dimensions; by individual piece or across sections of the portfolio). This finding led researchers to conclude that the state could not report on the percentage of students scoring at each point on each of the writing traits, nor could it provide comparative data across districts and schools (Koretz, Stecher, Klein, McCaffrey, & Deibert, 1993). Researchers suggested that inadequate rubrics, insufficient training of scorers, and lack of standardization of portfolio tasks most likely contributed to the lack of reliability.

In terms of validity, the Vermont results were "not persuasive" (Koretz, Stecher, Klein, & McCaffrey, 1994). The correlation between the portfolio scores and the Uniform Writing assessment were moderate, as one might expect from other research; however, these same levels of correlations were found between writing portfolios and a multiple choice math test. In addition, they found little difference between scores on papers students selected as "best pieces" and scores for the rest of the writing portfolio, a finding that

is inconsistent with other evidence suggesting lack of generalizability across different writing tasks (Dunbar, Koretz, Hoover, 1991). Validity was also brought into question in terms of portfolio implementation. In keeping with local autonomy, researchers found great variability in teachers' implementation of portfolios, resulting in a wide range of types of work included in the portfolios as well as a wide range of teacher support for the work, all of which would raise validity questions. Principals reported that, although the assessment system placed sizable demands on schools for resources—especially in the area of time and support—they thought it was a worthwhile burden (Koretz et al. 1993). Because they felt the burden fell primarily on the teachers, the majority of principals provided release time to help ease the stress.

In contrast to statewide initiatives on portfolios, several school districts have tried to implement literacy portfolios with the dual focus of accountability and improvement of instruction. Measurement researchers have studied both the ARTS PROPEL middle/high school writing portfolios in Pittsburgh and early literacy portfolios in Rochester, New York. Portfolios from Pittsburgh Public Schools (LeMahieu, Eresh, & Wallace, 1992) grew out of the ARTS PROPEL project, a privately-funded project to design instruction-based assessment in visual arts, music and imaginative writing. The writing portfolios were complied by students in grades 6–12 from a folder of their classroom writing. Using a set of guidelines, students selected four pieces (including drafts as well as final copies) and provided several written reflections on their writing processes, rationale for their selections, and the criteria they used for judging their work. As a result, there was less required commonality across portfolios than in the Vermont Portfolio Assessment. Portfolios were scored by a small group of highly trained district teachers and administrators using a rubric that reflected a decade-long district-wide history of professional development in writing. Judges assigned identical scores for 45 to 56% of the portfolios. Interrater correlations ranged from .80 to .84 across three scoring dimensions (accomplishment, processes/strategies, and growth). In addition, researchers found that portfolio scores were highly related to the classroom writing opportunities students were afforded. Students in classrooms judged to have teachers with an "intense" writing practice scored significantly better than those in classrooms judged to be moderately or not at all intense. Interestingly, portfolio scores were more strongly correlated with a standardized reading test than with a standardized direct writing measure.

The Rochester portfolios grew out of curriculum reform initiated three years before portfolios were sequentially implemented in the primary grades. The portfolios were designed by teachers to be scored by classroom teachers rather than by external scorers. They include both required pieces (e.g., writing samples, letter-sound assessments, observations, and anecdotal records) and optional pieces for reading and writing collected on a regular schedule. Teachers scored the portfolios and then assigned each child to a developmental stage specified in a "rubric." Supovitz, MacGowan, and Slattery (1997) compared the ratings given by Rochester classroom teachers and outside evaluators. They found interrater correlations between .58 and .77 with more consistency in reliability for writing than for reading. They found that external reviewers had difficulty scoring "thin" evidence found for reading both because few reading pieces were required in the portfolios and because teachers rarely included the required (or any additional) read-

ing evidence. When reading evidence was included, they found that outside raters had difficulty judging the work and applying judgments to developmental levels. Their findings also suggested that classroom teachers scored students significantly higher than outside raters in the area of reading, where the lack of portfolio evidence was most likely supplemented by teachers' knowledge. There were no significant differences across scorers in writing. In a second study, Supovitz and Brennan (1997) found that gender, socioeconomic, and racial inequities existed when portfolio performance was compared to standardized test performance, although the Rochester portfolios closed the gaps between blacks and whites and widened the gaps between boys and girls.

Questions about the variability in portfolio contents across students have been raised with respect to the influence on reliability, but the issue is pertinent with respect to validity as well. If students receive different levels of support or if evidence is simply not available, then judgments about students' abilities will be open to question. Two studies shed light on this point. In one, Herman, Gearhart, and Baker (1993) were able to get satisfactory levels of interrater agreement for portfolios containing only narrative and summary writing, but they discovered that students' scores were substantially different across different contexts (standard writing prompt vs. portfolio work; analytic vs. holistic rubrics; scoring of individual pieces vs. the total portfolio; narrative vs. summaries). In fact, two-thirds of the students classified as competent using portfolios scores were not judged competent on the standard writing assessment. This led the researchers to question the validity of portfolio scores and to look further behind the actual work. So, in another study, Gearhart, Herman, Baker, & Whittaker (1993) asked, "Whose work is it?" that is contained in students' writing portfolio. Teachers were asked to rate the level of instructional support for writing assignments in students' portfolios (grades 1-6). They found variability in the amount of support that teachers provided to students, the time students spent on assignments, and the extent to which work was copied. Furthermore, students received different levels of support depending upon whether they were low- or high-achieving, and teachers with more portfolio experience provided more teacher support. Not only was student work influenced by the level of teacher support, but this support was provided differentially across students and classrooms.

In an effort to look more closely at classroom-embedded performance assessment, Shepard and her colleagues (Shepard, Flexer, Hiebert, Marion, Mayfield, & Weston, 1996) examined the effects of a professional development project which was designed to help teachers use performance assessments as part of regular instruction in reading and mathematics. They reasoned that embedded performance assessments would improve learning by introducing challenging tasks that were consistent with curricular goals and by helping teachers clarify their understanding of their students, thereby informing their instruction. This study represents a shift from the other studies in this section in two important ways: (a) It integrates expertise in subject matter, teacher change and assessment in the design, implementation and analysis; and (b) it integrates professional development with a study of new assessments and student learning. Although the authors reported no gains in student learning in reading on the outcome measure (Maryland School Performance Assessment Program), they offered explanations that are consistent with other studies in both the policy and literacy sections. Specifi-

cally, they found that, although teachers were familiar with the district curricular framework before the project began, their motivation and instructional practices were not congruent with the framework. What the researchers conceived of as a professional project to introduce classroom performance assessment evolved into more of a project on literacy instruction and assessment. They concluded that performance assessments alone are not enough to improve teaching; long term professional development is also needed.

Summary and Implications

Measurement researchers tend to focus on the feasibility of new assessments from a technical perspective (reliability and validity) and on their desirability (consequential validity), often relying on statistical procedures and self-reports. Like policy researchers, measurement researchers generally have not distinguished among different subject areas in their targets for study or in their conclusions and recommendations, even though Linn (1998) (a prominent measurement researcher) has found differences in student performance across subject areas and within subscales of the same subject area. Overall, the measurement research on literacy assessment reform suggests that there is still uncertainly about the ability of performance assessments to provide reliable and valid data for accountability. As we might expect, contextual factors such as the nature of the classroom instruction and support provided to students are difficult to control across classrooms. Another source of difficulty is the degree of teacher involvement and teacher choice in the assessment. Several studies suggest that more standardization in the assessment artifacts, more expertise in the scorers, or more specificity in the outcomes could address problems of reliability and validity. However, these suggestions fly in the face of an important rationale for new standards and assessment—the professional development of teachers that is fostered through their involvement in the development and scoring of new assessments. As for the issue of consequential validity, data indicate that new assessments and standards have some influence on teachers' practices but that the accountability factor creates stress for teachers and raises questions about how well they implement new assessments. Whether we look at issues of feasibility or desirability of new assessments, the studies in this section highlight the tension between assessment for accountability and assessment for instructional improvement. They also raise questions about the quality and the influence of assessment as a policy tool (see Linn, 1998 and Mehrens, 1998 for more in-depth measurement perspectives across subject areas).

The Literacy Perspective

Literacy has been a centerpiece in efforts to push ambitious reforms in teaching and learning. National reports such as *A Nation at Risk* noted the failure of schools to provide the nation with a more literate populace as evi-

denced by allegedly declining verbal SAT scores and less than encouraging results of National Assessment of Educational Progress (NAEP) reading assessments. At the same time, the research community expressed concern about the skills-based conceptualizations that were guiding assessment and instruction in reading (see Curtis & Glaser, 1983; Guthrie & Kirsch, 1984; Linn, 1986); they were less concerned with writing because, as we noted previously, writing process and direct assessment of writing had gained popularity in the 1970s. These concerns led the National Academy of Education's Commission on Reading to issue the report entitled Becoming a Nation of Readers (BNR) (Anderson, Hiebert, Scott, & Wilkinson, 1985). The essence of BNR was that reading is a holistic, constructive process rather than the aggregate of a series of isolated subskills, and that curriculum, instruction, and assessment should reflect this view of reading. In many respects, BNR represented a subject-matter specific version of the national reports calling for more attention to higher-order thinking, and it became a conceptual framework for literacy researchers who were becoming involved in local, state, and national policy initiatives.

A constructivist view of reading was also evident in a number of state efforts to develop curriculum frameworks, objectives, and assessments in reading and language arts. For example, in 1984, Michigan put forward a "new" definition of reading as "the process of constructing meaning through the dynamic interaction among the reader, the text, and the context of the reading situation" (Wixson & Peters, 1984). This definition then served as the basis for new state Essential Goals and Objectives in Reading (1986). Given that BNR was written in Illinois, and Michigan was promoting a similar conceptualization of reading through its new definition, it is not surprising that these two states led the way in developing statewide reading assessments that better reflected constructivist reading theory and knowledge (Valencia, Pearson, Peters, & Wixson, 1989; Wixson, Peters, Weber, & Roeber, 1987).

Literacy researchers and state curriculum specialists worked with measurement specialists in Michigan and Illinois to develop a new generation of reading assessments consistent with new views of reading and new student outcomes (Peters, Wixson, Valencia, & Pearson, 1993). Although there was a precedent for this type of collaboration in the development of the NAEP tests, NAEP had little impact on the development of curriculum, instruction, and assessment at state, district, and school levels because it had been designed to provide information only at the national level. Rather, it was the large-scale reform efforts of the 1980s described previously in this chapter that brought together literacy researchers and curriculum and measurement specialists to effect the types of changes being called for by the research community, policymakers, and the public at large.

The constructivist perspective on reading being promoted in various national and state policy documents also influenced the development of a new Reading Framework for NAEP reading tests. The 1992 Reading Framework, which was also used in 1994 and 1998, indicates that "Reading for meaning involves a dynamic, complex interaction among three elements: the reader, the text, and the context" (NAGB, n.d., p. 10). Although NAEP continued to use direct writing assessment as it had done in the past, it began to include some open-ended reading items to address the new definition of reading and recommendations for new forms of reading assessment. The first year this framework was in effect was also the first year of the vol-

untary, trial program to administer NAEP in a way that allowed for state-by-state comparisons, and the year in which the NAEP special study on oral reading fluency was conducted (Pinnell, Pikulski, Wixson, Campbell, Gough, & Beatty, 1995). Both of these innovations came in response to demands by many that NAEP respond to the need for higher levels of accountability in general and specifically in the area of reading.

With the publication of the California English-Language Arts Framework in 1987, the treatment of reading and writing as separate subject areas began to give way to the idea of integrated language arts consisting of listening, speaking, reading, and writing. "Language arts became a discipline concerned with major universal themes, the human condition, exploring life experiences, and social agendas introduced through quality literature" (Gonzales & Grubb, 1997 p. 696). The California Framework pushed the definition of reading and language arts beyond a purely constructivist perspective to a more social-constructivist perspective with its emphasis on "transactions" as opposed to interactions with text and considerations of the sociocultural experiences students bring to text.

Prominent among the policy initiatives adopted by the California State Department of Education to support the framework was the California Learning Assessment System (CLAS). The CLAS reading assessment was designed to evaluate the success of the language arts curriculum, and to help districts and schools understand how well students were internalizing the strategies that encourage them to construct understandings beyond the school setting. CLAS took seriously the call for integration of the language arts by including open-ended written responses to reading selections, having students work collaboratively on some sections of the assessment, and tying some of the direct writing prompts to reading selections (Weiss, 1994).

In 1992, when the new CLAS assessments were being implemented, the contract to develop national English language arts standards was awarded by the U.S. Department of Education (DOE) to the Center for the Study of Reading at the University of Illinois, in collaboration with the International Reading Association (IRA) and the National Council of Teachers of English (NCTE). Before these standards were completed, however, the contract was terminated by the U.S. DOE for lack of satisfactory progress, reflecting differences in perceptions about what constitutes appropriate standards in English language arts. The project continued under the auspices of IRA and NCTE and concluded with the publication of the Standards for English Language Arts in 1996.

Consistent with the California Framework, the NCTE/IRA standards defined English language arts as listening, speaking, reading, writing, viewing, and representing. By the time the NCTE/IRA standards were published in 1996, there was widespread concern about the direction that constructivist and sociocultural views of teaching and learning were taking curriculum, instruction, and assessment in all subject areas including reading. In 1994, yielding to pressure from conservative groups, Governor Wilson vetoed legislation to continue funding for CLAS (Gonzales & Grubb, 1997), and the results of the 1992 and 1994 NAEP state-by-state comparisons placed California close to the bottom of the rankings in reading (Campbell, Donahue, Reese, & Phillips, 1996; Mullis, Campbell, & Farstrup, 1993). The "whole language" California framework was blamed for the failure of many California

children to learn to read, fueling a nationwide resurgence of the phonics vs. nonphonics "reading wars" which have surfaced every 10 or 20 years since the turn of the century (e.g. Chall, 1967; Flesch, 1955).

Most recently, there has been a shift away from attention on comprehension, writing, and integrated language arts to early reading, especially phonemic awareness and phonics. In 1998, some policymakers and educators promoted a return to skills-based definitions of reading that emphasized decoding as the only or primary concern of reading instruction. At the state level, there was a virtual firestorm of legislation focused on early reading that included new curriculum frameworks and standards, assessment mandates, textbook adoption guidelines, and mandates for teacher credentialing and professional development. For the first time since its inception in the mid-1970s, the U.S. DOE funded the national Center for the Improvement of Early Reading Achievement (CIERA) in 1997. At the same time, the National Research Council of the National Academy of Science commissioned a blue ribbon panel report on Preventing Reading Difficulties (National Research Council, 1998) and the National Institute for Child Health Development (NICHD) impaneled a group of experts to point educators and policymakers to the best research in reading instruction. With these recent events, national involvement in standards, assessment, and instructional strategies gained momentum.

The policy-oriented research we review in this section focuses on three areas related to literacy standards and assessments: on-demand reading and writing assessments, classroom-based assessments such as portfolios, and statewide language arts content standards. What distinguishes policy-oriented research conducted by literacy researchers from that conducted by policy and measurement researchers is a primary emphasis on the subject-matter content of standards and assessment and their validity and consistency with current literacy theory and research. Literacy researchers also focus on how policies shape classroom literacy practices, frequently gathering direct evidence of teaching and student learning as well as self-reported responses to policy. Literacy researchers tend to be less interested in the overall, or more systemic, effects of policy and reform.

Research

On-demand performance assessment.

Several literacy researchers have explored the influence of on-demand literacy assessments. In a series of studies, literacy researchers at the National Reading Research Center (NRRC) (Afflerbach, Almasi, Guthrie, & Schafer, 1996; Almasi, Afflerbach, Guthrie, & Schafer, 1995; Guthrie, Schafer, Afflerbach, & Almasi, 1994) investigated the effects of the Maryland State Performance Assessment Program (MSPAP), a multipronged reform effort that includes learning outcomes, a performance assessment, guidelines for school decision-making, and suggestions for staff development. Using semi-structured interviews, similar to those used by measurement researchers, they found that, one year after implementation, there was some limited understanding of the Maryland learning outcomes among country/district language arts administrators but no widespread consensus on the reading/language arts outcomes included in the MSPAP. Nevertheless, the administra-

tors believed that the performance assessment was moderately aligned with their local curriculum.

Although there was little to no reported change in school governance or teacher decision-making, administrators did report some change in instructional practices, including integrating reading and writing within content areas and the use of more trade books. In schools nominated as implementing positive innovations in response to the MSPAP, teachers and administrators reported changes in instructional tasks, methods, materials, and learning environments which reflected the nature of the MSPAP and learner outcomes in literacy. They also reported administrative support for change, including professional development, and a positive influence on students' motivation for reading and writing. The researchers identified several barriers to implementation even in the schools that were identified as successful implementors: lack of alignment between classroom instruction and assessment and the MSPAP assessment; insufficient resources such as time and money for professional development; testing logistics; and communication between the state and schools about the rationale and nature of the assessment program. They suggested that better communication and support is needed between state and local school districts if implementation is to be effective. Change, they argued, requires more than development of assessment materials and procedures.

Other researchers have examined more directly the relationship between new statewide writing assessments and classroom instruction. Two studies highlight the difficulties in achieving effective reciprocity between instructional practice and assessment. Goldberg, Roswell and Michaels (1995/1996) examined whether the MSPAP which required students to engage in the writing process (including drafting, peer response, revision, and writing a final draft), produced improved performance in writing. Specifically, researchers were interested in the extent to which students engaged in effective peer response, revision, and final drafts during testing. Using results from the MSPAP and observations from test taking in grades 3, 5, and 8, they found that students did not use revision or peer response to improve their final writing; their changes were minimal and focused on surface-level features, and their peer responses were unengaged. Goldberg et al. suggested that the constraints of large-scale assessment (i.e., assigned topics, limited and prescribed time blocks, use of revision and response worksheets, collaboration with assigned partners rather than classmates) may inhibit students' motivation and ability to engage in revision and peer response. They concluded that testing situations may not be able to mirror some aspects of good instructional practice.

Similarly, Loofbourrow (1994) conducted a case study of how two eighth-grade teachers interpreted and enacted the California Assessment Program (CAP) in writing into their classroom instruction. This study was conducted at a time when CAP was a high-stakes assessment and when it was not aligned with California curriculum guidelines for teaching writing. She found that, when there was misalignment between a high-stakes test and statewide recommendations for curriculum and instruction, teachers attended more to the form and content of the assessment. In this case, although middle-school teachers had students write across a wider variety of genres (an emphasis of both CAP and instructional recommendations), most of their writing assignments mirrored the test-like setting of CAP (e.g. lim-

ited time, one- to two-page writing assignments, teacher-assigned topics, focus on one of the eight CAP modes, emphasis on form over function). Many sound curricular and instructional recommendations were put aside as teachers attended to the specific form and content of the assessment.

Allington and his colleagues have taken a different approach to on-demand assessments, exploring the effects of on-demand assessment policies on special needs students and on the system as a whole. In several studies, Allington and McGill-Franzen (1992a, 1992b, McGill-Franzen & Allington, 1993) highlighted the changes in the incidence of retention, remediation, and identification of students as handicapped across a 10-year period when New York State increased high-stakes assessment and accountability. An increasing proportion of elementary children were retained or identified as handicapped in grades K-2, the grades that preceded the grade-3, high-stakes reading assessment. There was no corresponding trend for remediation. The researchers suggested that, although it was unlikely that the reform was intended to increase numbers of children retained or placed in special education, the net effect was that these low-achieving students were removed or delayed from the accountability stream. As a result, scores at the targeted grades were likely to rise without improved learning—the sample of students tested was simply limited. In fact, these researchers found that across all grades, schools that had been historically low-performing but seemed to be improving since the implementation of high-stakes assessment, had three times the number of students identified for special education or retained as compared with historically high-performing schools. Retention and identification, they argued, are expensive and ineffective ways to produce real gains.

Classroom-based assessment.

Most studies of classroom-based assessment literacy have investigated the effects of portfolio assessment; some have focused on statewide policies, and others on district-wide efforts. Two interesting lines of research have come from literacy researchers who have examined effects of statewide writing portfolio assessments in Vermont and Kentucky. Studies of the Vermont writing portfolio (Lipson, 1997, Lipson & Mosenthal, 1997; Mosenthal, Lipson, Mekkelsen, Daniels, & Jiron, 1996; Mosenthal, Mekkelsen, & Jiron, 1997) examined teachers' perspectives on the influence of the portfolio mandate and how they used the portfolios in classroom instruction and assessment. Unlike the research on Vermont portfolios cited in the previous assessment section, this work focused specifically on the writing portfolio and used a combination of surveys, interviews, and in-depth case studies and observations of 12 teachers. In addition, these researchers analyzed their findings in terms of teachers' different theoretical perspectives and beliefs about writing instruction.

Surveys were administered to fifth-grade teachers before and after the first year of implementation. The majority of teachers reported that their writing instruction had improved; they incorporated more writing into their classrooms and, once more writing was in place, they used the portfolio scoring criteria as part of their instructional talk with students. Although teachers embraced portfolios for instructional purposes, they did not seem to use portfolios or the criteria for assessing writing in the classroom, which suggested that they may not have a "shared standard" for student performance. Furthermore, teachers were strongly opposed to the scoring and public reporting of results.

Most interestingly, surveys and observations revealed that teachers with different beliefs about teaching writing changed in different ways. Those who already emphasized writing processes in their instruction felt most positively about the state assessment but changed very little because they had little need to change. In contrast, those teachers who were more dependent on curriculum and less child centered used portfolios for organizing writing, but did not integrate them into their teaching. Finally, those teachers who paid little attention to writing processes or portfolios were most negative and changed their practices very little.

The Kentucky writing portfolios have also been studied by literacy researchers. Bridge, Compton-Hall and Cantrell (1997) replicated a 1982 study to determine changes in the amount and type of writing elementary students were engaged in and the nature of writing instruction provided by teachers. They studied changes in one school district using written surveys of more than 200 teachers and classroom observations of teachers' instruction and writing activities of targeted students in 12 classrooms. Across both observations and surveys, they found a twofold increase in the amount of time students spent engaged in writing as compared to 1982; the biggest increase occurred at grade 1. This finding confirms results of other studies of reform in Kentucky (Bridge, 1994; Raths & Fanning, 1993).

Bridge et al. also looked closely at quality of the writing. They found a sizable increase in the amount of time spent on higher-level writing tasks such as crafting, and revising, and a decrease in the time students spent filling in words on worksheets or copying, which was dominant in 1982. In addition, teachers reported major changes in the way they responded to students' writing, shifting to greater use of teacher and group conferences and a decrease in assigning grades to students' writing. Teachers reported that, in large part, changes in their writing instruction could be attributed to the Kentucky assessments, although the authors acknowledged that most teachers were more knowledgeable about the writing process in 1995 than in 1982. Although more than 50 percent of teachers reported substantive changes in their writing instruction since the Kentucky Education Reform Act (KERA), about one-third of them reported little change because their instruction was already in line with the new assessments. Like the Vermont studies, this study highlights the increase in amount of writing and the differential impact of policy on teachers whose instruction is more or less aligned with new assessments.

A slightly different perspective on the Kentucky reform comes from a case study of nine high school English teachers faced with implementing the state portfolios in the second year of the mandate (Callahan, 1997). This study found that teachers had not yet received much professional development regarding the assessment and that they viewed the portfolio as a "test of their competence." Consequently, although the assessment did change the amount and kind of writing students did to fit with portfolio requirements and prompted teachers to internalize and use scoring criteria during instruction, teachers put their energy into "the visible, procedural elements of the assessment" rather than integrating it into their instruction. It remained a separate and intimidating burden to them. This lack of attention to professional development is also reported in studies by Gooden (1996) and by Miller, Hayes, and Atkinson (1997). Their work suggests that, in some states, having a high stakes assessment in place was assumed to be sufficient

to promote teacher change or to encourage local districts to provide support for change. Unfortunately, this did not happen.

Several literacy researchers have focused on classroom-based assessment at the district level (Hoffman, Worthy, Roser, McKool, Rutherford, & Strecker, 1996; Salinger & Chittenden, 1994; Valencia & Au, 1997). In all three of these studies, the researchers were interested in whether assessments could serve the dual purpose of improving instruction and providing accountability information. In all cases, the researchers worked directly with teachers and school districts in ongoing professional development activities focused on literacy curriculum and instruction as well as assessment implementation. Two studies focused early literacy assessment at the district level—the South Brunswick Early Literacy Portfolio (Salinger & Chittenden, 1994) and the Primary Assessment of Language Arts and Mathematics (PALM) (Hoffman, Worthy, Roser, McKool, Rutherford, & Strecker, 1996). The South Brunswick portfolio included specific content aimed at early literacy (K-2), and specific procedures and timelines for data collection. Using a developmental scale, teachers rated students on one component—strategies for making sense of and with print. The PALM model was somewhat different in that it combined three assessment elements: classroom-embedded assessments, a week-long on-demand assessment, and "taking a close look" assessments, which teachers used to gather additional information on particular students. The on-demand assessment and a developmental profile based on classroom-embedded and "take a closer look" information were scored. Using a combination of artifacts, interviews, and documents as data sources, Hoffman et al. and Salinger and Chittenden combined qualitative and quantitative analyses. In both studies, teachers reported that they could use assessment information for instructional purposes and that using these assessments was consistent with and enhanced their practice. Although teachers at both sites struggled with management and time issues, they all felt that the results justified the effort. Teachers viewed participation in professional development as critical to their success. Student data from both sites were able to be reliably scored, making the assessments useful for accountability purposes at a district level. In addition, statistical analysis of the PALM data revealed that all three components of PALM (classroom-embedded, on-demand, and taking a closer look) contributed significantly to the prediction of students' scores on a norm-referenced reading test.

The third example in this group of district-level efforts involved a cross-district study (Valencia & Au, 1997; Au & Valencia, 1997). This approach is different from others in that it addressed the question of whether common but not identical curriculum standards and portfolio structures could produce effective cross-site analysis and cross-site teacher learning. In addition, it examined the contextual factors that influenced portfolio implementation and used a portfolio model in which students chose a substantial number of pieces. Valencia and Au found that classroom portfolios contained artifacts consistent with the constructivist literacy curriculum frameworks at both sites. Although teachers were expected to include several required or "on-demand" pieces, some portfolios were missing needed evidence. This was a result of different emphases in different classrooms and the difficulty teachers had documenting particular aspects of reading. Yet once teachers were aware of what was missing, they were confident they could include it. Teachers reached a high level of agreement when rating portfolios from both sites, and they enhanced their knowledge of teaching, learning, and

assessment through the scoring process. Valencia and Au suggested that these results were a function of a supportive system for implementation which included district support, low-stakes, long-term professional development focused on assessment and instruction, and gradual implementation with an emphasis first on curriculum and instruction. They also suggested that the combination of required and optional pieces and the scoring process encouraged the flexibility of implementation and specificity of performance standards needed for portfolios to address both accountability and improved instruction.

Stephens, Pearson, Gilrane, Roe, Stallman, Shelton, Weinzierl, Rodriguez and Commeyras (1995) also addressed contextual influences in their study of the relationship between assessment and instruction. Using in-depth case studies of elementary schools in four school districts, they examined how decisions were made and how that process influenced the relationship between assessment and instruction. Qualitative analysis revealed that the relationship was not straightforward; the unique decision-making model in each district influenced the relationship. When the teachers had little authority or power over instructional decision-making, or when administrators were controlled by district staff, an "assessment-as-test" mentality drove instruction. In other words, when responsibility and accountability were to external forces, tests did drive instruction, and not necessarily in positive ways. When the culture of the district was one of responsibility to individual learners and decisions were based on individual or collective perspectives of teachers, assessment as test did not appear to drive instruction. Stephens et al. raise the question of whether reform aimed at teacher empowerment can coexist with external accountability when school culture exerts such a strong influence on teachers' practice.

Language arts standards.

Few literacy researchers have conducted research aimed directly at either state or national English language arts and, as we have noted previously, policy researchers interested in standards often analyze them in terms of larger reforms efforts, without specific regard for subject area. Three types of studies characterize the nature of standards research from a literacy perspective: document analysis, study of teachers' practices, and the alignment of standards with assessment.

Wixson and Dutro (in press) conducted a document analysis of 42 state standards in early reading/language arts as a way to gauge how the variability in standards might influence their translation into local curriculum, instruction and assessment. They found that the majority of state documents did not provide specific benchmarks or outcomes at grades K-3, that the documents varied in the way they conceptualized and organized the area of reading, and that many of them included inappropriate content and/or ignored important content. When documents did provide benchmarks, many did not provide a logical developmental progression across grades, and many of the benchmarks themselves were overly specific or overly broad. In the former case, Wixson and Dutro concluded, districts are provided insufficient guidance; in the latter, the curriculum becomes prescriptive without much flexibility for local interpretation. They recommend the need for balance between specificity and generality if standards are to help local educators engage in conversations needed to advance teaching and learning.

In a different approach to examining standards, McGill-Franzen and Ward (1997) first reviewed documents to determine the fit between New York State Language Arts and Social Studies frameworks with national standards. They then conducted case study interviews with K-4 teachers in four districts to determine how state standards were incorporated into teachers' practices. All the participants in their case studies had been involved in some sort of school-wide language arts curriculum development projects aimed at helping teachers reconceptualize teaching. Consistent with Wixson and Dutro's (in press) recommendations, they found New York State standards reflected the national standards in orientation to reading process and learning, and actually went beyond national standards to provide a level of specificity that helps teachers know what students should know and be able to do at different developmental levels. However, they also found that teachers interpreted the state standards differentially. If teachers were under pressure to improve scores on tests (which were not aligned with the standards), and worked under conditions that restricted their authority and responsibility for instructional decision-making, they were less likely to reconceptualize curriculum and evaluation in their schools. Since this study, New York State has restructured their assessments to align with their curriculum. We do not yet have data to know whether the results of McGill-Franzen & Ward (1997) will be replicated with the new assessments.

The last approach to research on standards is found in a study by Bruce, Osborn, and Commeyras (1993) in which they examined the alignment between NAEP reading assessment items and the NAEP reading framework (standards). Using data from interviews, expert panels, and surveys of hundreds of literacy educators, Bruce et al. concluded that although most literacy experts agreed that the NAEP framework reflected current research and practice, the experts judged the alignment between framework and test items to be "murky." Items could not be mapped clearly onto the framework and, in practice, the items often failed to capture the intent behind the framework. Even with a sound framework, the translation into large-scale assessment items was problematic.

Summary and Implications

Literacy researchers bring a deliberate subject-matter focus to bear on questions of standards and assessment. For the most part, they look more deeply at literacy than either policy or measurement researchers by examining specific aspects of literacy instruction (e.g., writing process, qualities of writing, alignment of assessment with constructivist curriculum frameworks in literacy, specificity of state standards) and by situating much of their work in classrooms or in direct interactions with teachers. The studies in this section suggest that instructional change in language arts does occur with reform, but that it is mediated by teachers' beliefs, knowledge, and their sense of accountability pressure. In studies which integrate professional development with assessment reform, results are most positive both in terms of teachers' learning and attitudes toward change, and in terms of usable assessment information. Reform without this support seems to produce surface-level change and questionable assessment practices. At the same time, the work on standards and implementation of new assessments suggests that the translation from literacy research to standards, and from standards to

assessment, is not straightforward. Overall, an emerging theme is one of tension between the need for accountability and specificity on the one hand, and teacher decision-making and flexibility in interpretation on the other.

Conclusions and Future Directions

Throughout this report we have attended to both the nature of policy-oriented research on literacy standards and assessments and the impact of standards and assessment on literacy practice and student learning. Our conclusions address both of these issues.

With regard to the nature of policy-oriented research, it is clear that the research differs with regard to questions, methods, and audience as a function of the perspective from which it arises. In general, we see policy researchers concerned with broad reforms involving standards, assessments, reorganization, governance, and the like; literacy is simply one of the subjects, and standards and assessment are two of the "tools" or levers of reform they study. Policy researchers' questions focus on the system, and their data are gathered through teachers' reported and actual practices. For the most part, we found few policy researchers distinguishing among subject areas within policy or spending extended time in literacy classrooms. Measurement researchers, as we might expect, are most interested in the assessment components of reform and are particularly concerned with validity issues and with the psychometric qualities of new assessments that are needed for accountability and policy purposes. For the most part, they rely on statistical analysis and, to some extent, self-reports, interviews, and artifacts to address their questions. Literacy researchers generally ask questions about instruction and learning in relation to research and theory. Do new standards and assessments result in better reading and writing instruction? Do they advance teacher understanding? Are the reforms consistent with sound research and theory on literacy learning? Just as literacy is the vehicle for many policy studies, policy is the vehicle for many literacy studies. Literacy researchers typically look closely at actual classroom practices, teachers' understanding, and artifacts of students literacy learning, and they work more directly with teachers than either policy or measurement researchers. For the most part, their new assessments are not subjected to the rigor of measurement researchers' criteria, and the policy contexts for their work are not considered in a systemic way.

The picture that emerges is that of a trade-off between general and in-depth information. Studies that address general questions provide information that is useful for understanding the larger issues of systemic reform (e.g., restructuring, governance, standards, and assessment) and the contexts in which these reforms are implemented. In contrast, research that addresses questions about classroom practice in relation to specific subject matter provides insight into what happens at the individual teacher and student levels. It attends to teacher understanding and practice, and student learning, often without specific attention to the policy environment in which change is enacted. As policy-oriented research grows and matures, we see a greater need to attend to both the macro- and microview of reform, practice, and

learning, and we suspect there will be more cross-over among the studies representing policy, measurement, and literacy perspectives.

With respect to the influence of policy, it is clear that literacy standards and assessment do have an influence on teachers' beliefs and practice. However, the influence is not always in the expected or desired direction. The effect is mediated by numerous factors: teachers' knowledge, beliefs, and existing practices; the economic, social, philosophical, and political conditions of the school or district; the stakes attached to the policy; and the quality of the support and lines of communication provided to teachers and administrators. It is equally clear that policy alone is not sufficient to promote change; simply implementing new assessments or creating new standards does not ensure improved teaching or learning. What is less clear, however, is what it would take to promote change in the desired direction or to ensure improved teaching and learning. To be sure, discipline-specific professional development is implicated in many studies, but we need to know more about professional development processes and the quality of those processes. How, for example, do teachers and districts learn about new literacy standards and assessments? How are districts and teachers supported to understand the theory and research that underlie new assessments or new content standards? How do these experiences shape teachers' understanding and practices? What are effective models for professional development? These are not only important questions for educators; they are critical to policymakers who are being asked to support professional development as part of reform (Elliott, 1996; Hart, 1996).

Among the factors mediating the effects on literacy teaching and learning, the research suggests that more specificity in standards and assessment promotes changes in the desired direction. A caution we would offer in this regard, however, is that many of the deeper levels of change in teacher beliefs and practices associated with literacy learning do not lend themselves to simple directives. There is a very fine line between offering sufficient guidance for teachers and districts to undertake substantive change, and being prescriptive in ways that work against teacher learning, decision-making, and flexibility. Similarly, the influence of standards and assessments is likely mediated by teachers' and administrators' stance toward policy. Do they see policy as a means for monitoring, controlling, or helping educators do their work? Why, for example, would teachers bring impoverished understandings of assessments and standards to policy work in their districts, yet demonstrate deep understanding in their classrooms? How do the messages teachers receive about standards and assessments fit or conflict with other policies in their environments such as mandated curriculum and materials, alternative teacher certification programs, site-based decision making, and the like?

Finally, there is still uncertainty about the quality of the new literacy assessments and standards; they must stand the test of scrutiny of policymakers outside education, as well as educators themselves. If the tools themselves are problematic, political credibility and deeper-order change are highly improbable. That said, we also suggest that although there is currently a move away from the more elaborate forms of performance assessments (i.e. California, Arizona, Kentucky), these decisions are rarely based on psychometric qualities alone. Policy, resources, and politics weigh heavily in deci-

sions about the feasibility and desirability of new assessments and new standards.

We conclude from our synthesis that there is a pressing need to conduct research on policy issues, such as standards and assessments, with specific attention to subject matter. There has been an almost tacit assumption among policy and measurement researchers that whatever holds true for one subject will likely be the case for others. When it is possible to look across subject areas such as mathematics and literacy, the analyses are rarely done. However, attention to depth of understanding of literacy processes, learning, and instruction is even more important. At the heart of this issue are questions about what it means to read and write with understanding; what teaching for understanding looks like in different classrooms; and what constitutes the domain of the English language arts curriculum. Future research must look both across and within the subject matter of literacy; this requires subject-matter expertise, and it requires more than self-report data.

Issues of student achievement will need to be confronted as well. Few studies we reviewed included direct measures of student learning. In one sense, this is understandable since most reform is fairly new and change is a long-term process. Nevertheless, pressure is mounting on educators to show results in terms of achievement. Future researchers will need to address the challenge, finding meaningful ways to document student achievement while documenting formative measures of progress such as parents' understanding of instructional goals, teachers' priorities and their practice, teacher understanding, and surface-level changes in materials and activities.

As we write this report in the late 1990s, there continues to be a groundswell of new policies related to literacy standards and assessment, and there is new interest in policies related to instructional strategies. Whether we like it or not, literacy researchers have been drawn into policy. At worst we will be recipients of policy; at best we will be informers of policy. In our opinion, the best way to influence policy and teacher development is for policy, measurement, and literacy researchers to work together and to communicate the findings of their collaborative work in a wide range of journals and reports, and through participation in state and national councils. Literacy researchers need to become knowledgeable about policy research and about the policy contexts in which their research is conducted. At the same time, we must reach out to policy researchers and measurement researchers, bringing to their work a deep understanding of the subject matter of literacy and the pedgogical content knowledge needed to teach well. Without this collaborative commitment, policy will not reflect or inform meaningful changes in literacy teaching and learning; measurement will not encourage substantive instructional change or provide useful assessment information to literacy educators; and literacy educators will not have a voice in policy and measurement arenas. With a collaborative research agenda and a wider audience, we can improve the lives of children and teachers.

References

Afflerbach, P. P., Almasi, J., Guthrie, J. T. & Schafer, W. (1996). ***Barriers to implementation of a statewide performance program: School personnel perspectives*** (Reading Research Report No. 51). Athens, GA: National Reading Research Center.

Allington, R. L., & McGill-Franzen, A. (1992a). Unintended effects of educational reform in New York. *Educational Policy, 6,* 397-414.

Allington, R. L., & McGill-Franzen, A. (1992b). Does high stakes testing improve school effectiveness? *Spectrum, 10,* 3-12.

Almasi, J. F, Afflerbach, P. P., Guthrie, J. T., & Schafer, W. D. (1995). ***Effects of a statewide performance assessment program on classroom instructional practice in literacy*** (Reading Research Report No. 32). Athens, GA: National Reading Research Center.

Anderson, R. C., Hiebert, E. H., Scott, J. A., & Wilkinson, A. G. (1985). ***Becoming a nation of readers: The report of the commission on reading.*** Washington, DC: The National Institute of Education.

Aschbacher, P. R. (1994). Helping educators to develop and use alternative assessments: Barriers and facilitators. *Educational Policy, 8,* 202-223.

Au, K. H., & Valencia, S. W. (1997). The complexities of portfolio assessment. In D. Hansen & N. Burbules (Eds.), *Teaching and its predicaments* (pp. 123-144). Boulder, CO: Westview.

Ball, D. L., & Cohen, D. K. (April, 1995). ***What does the educational system bring to learning a new pedagogy of reading or mathematics?*** Paper presented at the annual American Educational Reading Association conference, San Francisco.

Bridge, C. A. (1994). Implementing large scale change in literacy instruction. In Kinzer, C. & Leu, D. (Eds.), ***Multidimensional aspects of literacy research, theory, and practice.*** (Forty-third yearbook of the National Reading Conference, pp. 257-265). Chicago: National Reading Conference.

Bridge, C. A., Compton-Hall, M., & Cantrell, S. C. (1997). Classroom writing practices revisited: The effects of statewide reform on writing instruction. *Elementary School Journal, 98,* 151-170.

Bruce, B., Osborn, J., & Commeyras, M. (1993). Contention and consensus: The development of the 1992 National Assessment of Educational Progress in reading. *Educational Assessment, 1,* 225-254.

California Department of Education (1987). ***English-language arts framework for California public schools.*** Sacramento, CA: Author.

Callahan, S. (1997). Tests worth taking?: Using portfolios for accountability in Kentucky. *Research in the Teaching of English, 31 (3),* 295-336.

Campbell, J. R., Donahue, P. L, Reese, C. M., & Phillips, G. W. (1996). *NAEP 1994 reading report card for the nation and the states.* Washington, DC: Government Printing Office.

Chall, J. (1967). *Learning to read: The great debate.* New York: McGraw-Hill.

Cohen, D. K. (1995). What is the system in systemic reform? *Educational Researcher, 24(9),* 11-17.

Cohen, D. K., & Spillane, J. (1992). Policy and practice: The relations between governance and instruction. In G. Grant (Ed.), *The Review of Research in Education* (pp. 3-49). Washington, DC: American Educational Research Association.

Cuban, L. (1990). Reforming again, again, and again. *Educational Researcher, 19(1),* 3-13.

Curtis, M. E., & Glaser, R. (1983). Reading theory and the assessment of reading achievement. *Journal of Educational Measurement, 20,* 133-148.

Darling-Hammond, L. (1990). Instructional policy into practice: "The power of the bottom over the top." *Educational Evaluation and Policy Analysis, 12,* 233-241.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4,* 289-303.

Elliott, E. J. (1996). Literacy: From policy to practice. *Journal of Literacy Research, 28,* 590-595.

Elmore, R. F. (1990). Introduction: On changing the structure of public schools. In R. F. Elmore (Ed.), *Restructuring schools: The next generation of educational reform* (pp. 1-28). San Francisco: Jossey-Bass.

Firestone, W. A., Fuhrman, S. H., & Kirst, M. W. (1989). *The progress of reform: An appraisal of state education initiatives.* New Brunswick, NJ: Consortium for Policy Research in Education, Rutgers University.

Flesch, R. (1955). *Why Johnny can't read.* New York: Harper & Brothers.

Freedman, S. W. (1993). Linking large-scale testing and classroom portfolio assessments of student writing. *Educational Assessment, 1,* 27-52.

Fullan, M. G., & Miles, M. B. (1992). Getting reform right: What works and what doesn't. *Phi Delta Kappan,* 744-752.

Gearhart, M., Herman, J. L., Baker, E. L., & Whittaker, A. K. (1993). *Whose work is it? A question for the validity of large-scale portfolio assessment* (CSE Technical Report 363). Los Angeles: National Center for the Research on Evaluation, Standards, and Student Testing, University of California.

Goertz, M. E., Floden, R. E., & O'Day, J. (1995). *Studies of education reform: systemic reform (Vol. 1): Findings and conclusions.* New Brunswick, NJ: Consortium for Policy Research in Education, Rutgers University.

Goldberg, G. L., Roswell, B. S., & Michaels, H. (1995/1996). Can assessment mirror instruction? A look at peer response and revision in a large-scale writing test. *Educational Assessment, 3(4),* 287-314.

Gonzales, P. C., & Grubb, M. (1997). California's literature-based curriculum and the California literature project. In J. Flood, S. B. Heath, & D. Lapp (Eds.), ***Handbook for literacy educators: Research on teaching and communicative and visual arts*** (pp. 695703). Newark, DE: International Reading Association.

Gooden, S. (1996). A comparison of writing assessment portfolios in two states: Implications for large-scale writing assessment. In D. J. Leu, C. K. Kinzer, & K. A. Hinchman (Eds.), ***Literacies for the 21st century: Research and practice.*** (Forty-fifth yearbook of the National Reading Conference, pp. 88-99). Chicago: National Reading Conference, Inc.

Guthrie, J. T., & Kirsch, I. (1984). The emergent perspective on literacy. *Phi Delta Kappan, 65,* 351-355.

Guthrie, J. T., Schafer, W. D., Afflerbach, P. P., & Almasi, J. F. (1994). ***Systemic reform of literacy education: State and district-level policy changes in Maryland*** (Reading Research Report 27). Athens, GA: National Reading Research Center.

Hart, G. K. (1996). A policymaker's response. *Journal of Literacy Research, 28,* 596-601.

Haney, W. (1991). We must take care: Fitting assessments to functions. In V. Perrone (Ed.), ***Expanding student assessment*** (pp. 142-163). Alexandria, VA: Association for Supervision and Curriculum Development.

Herman, J. L. (1991). Research in cognition and learning: Implications for achievement testing practice. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 154-165). Englewood Cliffs, NJ: Prentice Hall.

Herman, J. L., Gearhart, M., & Baker, E. L. (1993). Assessing writing portfolios: Issues in the validity and meaning of scores. *Educational Assessment, 1,* 201-224.

Herman, J. L., & Golan, S. (1993). The effects of standardized testing on teaching and schools. *Educational Measurement: Issues and Practice, 12(4),* 20-25, 41-42.

Hieronymus, A. N., Hoover, H. D., Cantor, N. K., & Oberley, K. R. (1987). ***Writing supplement teacher's guide: Iowa Tests of Basic Skills.*** Chicago: The Riverside Publishing Company.

Hoffman, J. V., Worthy, J., Roser, N. L., McKool, S. S., Rutherford, W. L., & Strecker, S. (1996). Performance assessment in first grade classrooms: The PALM model. In D. J. Leu, C. K. Kinzer, & K. A. Hinchman (Eds.), *Literacies for the 21st century: Research and practice.* (Forty-fifth yearbook of the National Reading Conference, pp. 100-112). Chicago: National Reading Conference.

Jennings, N. E. (1995). ***Interpreting policy in real classrooms: Case studies of state reform and teacher practice.*** New York: Teachers College Press.

Koretz, D. M., Barron, S. Mitchell, K. J., Stecher, B. M. (1996). *Perceived effects of the Kentucky Instructional Results Information System (KIRIS).* Santa Monica, CA: RAND.

Koretz, D., McCaffrey, D., Klein, S., Bell, R., & Stecher, B. (1993). *The reliability of scores from the 1992 Vermont Portfolio Assessment Program.* (CSE Technical Report 355). Los Angeles: RAND Institute on Education and Training, National Center for Research on Evaluation, Standards, and Student Testing, University of California.

Koretz, D., Mitchell, K., Barron, S., & Keith, S. (1996). *Final report: Perceived effects of the Maryland School Performance Assessment Program* (CSE Technical Report No. 409). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, University of California.

Koretz, D., Stecher, B., & Deibert, E. (1992). *The Vermont Portfolio Assessment Program: Interim report on implementation and impact, 1991-92 school year* (CSE Technical Report No. 350). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, University of California.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont Portfolio Assessment Program: Findings and implications. *Educational Measurement: Issues and Practice 13(3),* 5-16.

Koretz, D., Stecher, B., Klein, S., McCaffrey, D., & Deibert, E. (1993). *Can portfolios assess student performance and influence instruction? The 1991-1992 Vermont experience* (CSE Technical Reporp No. 371). Los Angeles: National Center for the Research on Evaluation, Standards, and Student Testing, University of California.

LeMahieu, P. G., Eresh, J. T., & Wallace, R. C. (1992). Using student portfolios for a public accounting. *The School Administrator, 49(11),* 8-15.

Lindle, J. C., Petrosko, J., & Pankratz, R. (May 1997). *1996 review of research on the Kentucky education reform act.* Frankfort, KY: The Kentucky Institute for Education Research.

Linn, R. L. (1986). Educational testing and assessment: Research needs and policy issues. *American Psychologist, 41,* 1153-1160.

Linn, R. L. (1998). *Assessments and accountability.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20(8),* 15-21.

Linn, R. L., Grau, E., & Sanders, N. M. (1990). Comparing state and district test results to national norms: The validity of claims that 'Everyone is above average'. *Educational Measurement: Issues and Practice, 9(3),* 5-14.

Lipson, M. Y. (1997). *Teacher diversity as an influence on change: Capturing the multiple dimensions of teacher beliefs.* Paper presented at the

annual meeting of the American Educational Research Association, Chicago, IL.

Lipson, M. Y., & Mosenthal, J. (1997). *The differential impact of Vermont's writing portfolio assessment on classroom instruction.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Loofbourrow, P. T. (1994). Composition in the context of the CAP: A case study of the interplay between composition assessment and classrooms. *Educational Assessment, 2,* 7-49.

Lucas, C. K. (1988a). Toward ecological evaluation. *The Quarterly of the National Writing Project and the Center for the Study of Writing, 10(1),* 1-3, 12-17.

Lucas, C. K. (1988b). Recontextualizing literacy assessment. *The Quarterly of the National Writing Project and the Center for the Study of Writing, 10(2),* 4-10.

Lusi, S. F. (1997). *The role of state departments of education in complex school reform.* New York:Teachers College Press.

Massell, D., Kirst, M., & Hoppe, M. (1997). *Persistence and change: Standards-based reform in nine states.* Philadelphia, PA: Consortium for Policy Research in Education, University of Pennsylvania.

McDonnell, L. (February 1997). *The politics of state testing: Implementing new student assessments.* Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, University of California.

McDonnell, L., & Choisser, C. (September 1997). *Testing and teaching: Local implementation of new state assessments.* Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, University of California.

McGill-Franzen, A., & Allington, R. L. (1993). Flunk'em or get them classified: The contamination of primary grade accountability data. *Educational Researcher, 22(1),* 19-22.

McGill-Franzen, A., & Ward, N. (1997). *Teachers' use of new standards, frameworks, and assessments for English Language Arts and Social Studies: Local cases of New York State primary grade teachers.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

McLaughlin, M. W. (1987). Learning from experience: Lessons from policy implementation. *Educational Evaluation and Policy Analysis, 9,* 171-178.

Mehrens, W. A. (1998). Consequences of assessment: What is the evidence? *Education Policy Analysis Archives* [On-line serial], 6(13). Available http://olam.ed.asu.edu/epaa/.

Mehrens, W. A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice, 11(1),* 3-9, 20.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.), (pp. 13-103). New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher 23(2)*, 13-23.

Michigan State Board of Education. (1986). *Michigan essential goals and objectives for reading education.* Lansing: Author.

Miller, M. D., & Legg, S. M. (1993). Alternative assessment in a high-stakes environment. *Educational Measurement: Issues and Practice, 12(2)*, 9-15.

Miller, S. D., Hayes, C. T., & Atkinson, T. S. (1997). State officials' efforts to improve students' reading and language arts achievement with their newly designed end-of-grade assessments. In C. K. Kinzer, K. A. Hinchman, & D. J. Leu (Eds.), *Inquiries in literacy theory and practice*, (Forty-sixth yearbook of the National Reading Conference, pp. 91-100). Chicago, IL: National Reading Conference.

Mosenthal, J., Lipson, M. Y., Mekkelsen, J., Daniels, P., & Jiron, H. W. (1996). The meaning and use of portfolios in different literacy contexts: Making sense of the Vermont Assessment Program. In D. J. Leu, C. K. Kinzer, & K. A. Hinchman (Eds.), *Literacies for the 21st century*, (Forty-fifth yearbook of the National Reading Conference, pp. 113-123). Chicago, IL: National Reading Conference.

Mosenthal, J. H., Mekkelsen, J. E., & Jiron, H. W. (1997). *Agents of their own instruction: The teacher's perspective on the influence of the Vermont Assessment Program.* Paper presented at the annual meeting of the American Educational Association, Chicago, IL.

Moss, P. (1994). Can there be validity without reliability. *Educational Researcher, 23(2)*, 5-12.

Mullis, I. V. S., Campbell, J. R., & Farstrup, A. E. (1993). *NAEP 1992 reading report card for the nation and the states.* Washington, DC: U.S. Government Printing Office.

National Assessment Governing Board. (n.d.). *Reading framework for the National Assessment of Educational Progress, 1992-98.* Washington, DC: Author.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform.* Washington, DC: U.S. Government Printing Office.

National Council on Education Standards and Testing. (1992). *Raising standards for American Education.* Washington, DC: U.S. Government Printing Office.

National Educational Goals Panel. (1991). *National educational goals report: Building a nation of learners.* Washington, DC: U.S. Government Printing Office.

National Research Council. (1998). *Preventing reading difficulties in young children.* Washington DC: National Academy Press.

Nolen, S. B., Haladyna, T. M., & Haas, N. S. (1992). Uses and abuses of achievement test scores. *Educational Measurement: Issues and Practice, 11(2),* 9-15.

Peters, C. W., Wixson, K. K., Valencia, S. W., & Pearson, P. D. (1993). Changing statewide reading assessment: A case study of Michigan and Illinois. In B. R. Gifford (Ed.), *Policy Perspectives on Educational Testing* (pp. 295-391). Boston: Kluwer Academic Publishers.

Pinnell, G. S., Pikulski, J. J., Wixson, K. K., Campbell, J. R., Gough, P. B., & Beatty, A. S. (1995). *Listening to children read aloud.* Washington, DC: U.S. Government Printing Office.

Popham, J. W., Cruse, K. L., Rankin, S. C., Sandifer, P. D., & Williams, P. L. (1985). Measurement-driven instruction: It's on the road. *Phi Delta Kappan, 66,* 628-634.

Raths, J. & Fanning, J. (1993). *Primary school reform in Kentucky revisited.* Lexington, KY: Prichard Committee for Academic Excellence.

Resnick, L. B., & Resnick, D. L. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Future assessments: Changing views of aptitude, achievement, and instruction* (pp. 37-75). Boston: Kluwer.

Resnick, L. B., Resnick, D. L., & DeStefano, L. (1993). *Cross-scorer and cross-method comparability and distribution of judgments of student math, reading and writing performance: Results from the New Standards Project Big Sky Scoring Conference.* Los Angles: Center for Research on Evaluation, Standards and Student Testing, University of California.

Roeber, E., Bond, L. A., & Braskamp, D. (1997). *Trends in statewide student assessment programs, 1997.* Oak Brook, IL: North Central Regional Educational Laboratory and Council of Chief State School Officers.

Salinger, T., & Chittenden, E. (1994). Analysis of an early literacy portfolio: Consequences for instruction. *Language Arts, 71, 6,* 446-452.

Shepard, L. A. (1989). Why we need better assessments. *Educational Leadership, 46(7),* 4-9.

Shepard, L. A. (1991). Will national tests improve student learning? *Phi Delta Kappan, 72,* 232-238.

Shepard, L. A., & Dougherty, K. C. (1991). *Effects of high-stakes testing on instruction.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Shepard, L. A., Flexer, R. J., Hiebert, E. H., Marion, S. F., Mayfield, V., & Weston, J. T. (1996). Effects of introducing classroom performance assessments on student learning. *Educational Measurement: Issues and Practice, 15*(3), 7-18.

Simmons, W., & Resnick, L. (1993). Assessment as the catalyst of school reform. *Educational Leadership, 50(5)*, 11-16.

Smith, M. L. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher, 20 (5)*, 8-11.

Smith, M. L., Edelsky, C., Draper, K., Rottenberg, C., & Cherland, M. (1990). *The role of testing in elementary schools* (CSE Technical Report No. 321). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing, University of California.

Smith, M. L., Noble, A. J., Heinecke, W., Seck, M., Parish, C., Cabay, M., Junker, S. C., Haag, S., Taylor, K., Safran, Y., Penley, Y., & Bradshaw, A. (1997). *Reforming schools by reforming assessment: Consequences of the Arizona Student Assessment Program (ASAP): Equity and teacher capacity building* (CSE Technical Report No. 425). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing, University of California.

Smith, M. S., & O'Day, J. (1991). Systemic school reform. In S. Furhman (Ed.), *The politics of curriculum and testing: The 1990 yearbook of the Politics of Education Associations* (pp. 233-267). Philadelphia: Falmer Press.

Spillane, J. P. (1996). School districts matter: Local educational authorities and state instructional policy. *Educational Policy, 10*, 63-87.

Spillane, J. P. (1998). State policy and the non-monolithic nature of the local school district: Organizational and professional considerations. *American Educational Research Journal, 35*, 33-63.

Spillane, J. P. & Jennings, N. E. (1997). Aligned instructional policy and ambitious pedagogy: Exploring instructional reform from the classroom perspective. *Teachers College Record, 98*, 449-481.

Standerford, N. S. (1997). Reforming reading instruction on multiple levels: Interrelations and disconnections across the state, district, and classroom levels. *Educational Policy, 11*, 58-91.

Stephens, D., Pearson, P. D., Gilrane, C., Roe, M., Stallman, A. C., Shelton, J., Weinzierl, J., Rodriguez, A., & Commeyras, M. (1995). Assessment and decision making in schools: A cross-site analysis. *Reading Research Quarterly, 30*, 478-499.

Supovitz, J. A., & Brennan, R. T. (1997). Mirror, mirror on the wall, which is the fairest test of all? An examination of the equitability of portfolio assessment relative to standardized tests. *Harvard Educational Review, 67*, 472-506.

Supovitz, J. A., MacGowan III, A., & Slattery, J. (1997). Assessing agreement: An examination of the interrater reliability of portfolio assessment in Rochester, New York. *Educational Assessment, 4*, 237-259.

Valencia, S. W. (1991). Portfolios: Panacea or Pandora's Box? In F. Finch (Eds.), *Educational performance testing* (pp. 33-46). Chicago, IL: Riverside Publishers.

Valencia, S. W., & Au, K. H., (1997). Portfolios across educational contexts: Issues of evaluation, professional development, and system validity. *Educational Assessment, 4,* 1-35.

Valencia, S. W., Pearson, P. D., Peters, C. W., & Wixson, K. K. (1989). Theory and practice in statewide reading assessment: Closing the gap. *Educational Leadership, 47*(7), 57-63.

Weiss, B. (1994). California's new English-Language Arts assessment. In S. W. Valencia, E. H. Hiebert, & P. P. Afflerbacj (Eds.), ***Authentic reading assessment: Practices and possibilities.*** Newark, DE: International Reading Association.

Welch, C. (1991). ***Estimating the reliability of a direct measure of writing through generalizability theory.*** Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Wiggins, G. P. (1993). ***Assessing student performance.*** New York: Jossey-Bass.

Wixson, K. K., & Dutro, E. (in press). ***Standards for primary-grade reading: An analysis of state frameworks.*** Elementary School Journal. (Also available as CIERA Report 3-001. Ann Arbor, MI: Center for the Improvement of Early Reading Achievement.)

Wixson, K. K., & Peters, C. W. (1984). Reading redefined: A Michigan Reading Association position paper. *Michigan Reading Journal, 17,* 4-7.

Wixson, K. K., Peters, C. W., Weber, E. M., & Roeber, E. (1987). New directions in statewide reading assessment. *The Reading Teacher, 40,* 749-754.

About CIERA

The Center for the Improvement of Early Reading Achievement (CIERA) is the national center for research on early reading and represents a consortium of educators in five universities (University of Michigan, University of Virginia, and Michigan State University with University of Southern California and University of Minnesota), teacher educators, teachers, publishers of texts, tests, and technology, professional organizations, and schools and school districts across the United States. CIERA is supported under the Educational Research and Development Centers Program, PR/Award Number R305R70004, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

Mission. CIERA's mission is to improve the reading achievement of America's children by generating and disseminating theoretical, empirical, and practical solutions to persistent problems in the learning and teaching of beginning reading.

CIERA Research Model

The model that underlies CIERA's efforts acknowledges many influences on children's reading acquisition. The multiple influences on children's early reading acquisition can be represented in three successive layers, each yielding an area of inquiry of the CIERA scope of work. These three areas of inquiry each present a set of persistent problems in the learning and teaching of beginning reading:

CIERA INQUIRY 1
Readers and Texts

Characteristics of readers and texts and their relationship to early reading achievement. What are the characteristics of readers and texts that have the greatest influence on early success in reading? How can children's existing knowledge and classroom environments enhance the factors that make for success?

CIERA INQUIRY 2
Home and School

Home and school effects on early reading achievment. How do the contexts of homes, communities, classrooms, and schools support high levels of reading achievement among primary-level children? How can these contexts be enhanced to ensure high levels of reading achievement for all children?

CIERA INQUIRY 3
Policy and Profession

Policy and professional effects on early reading achievement. How can new teachers be initiated into the profession and experienced teachers be provided with the knowledge and dispositions to teach young children to read well? How do policies at all levels support or detract from providing all children with access to high levels of reading instruction?

www.ciera.org

CIERA
Center for the Improvement of Early Reading Achievement

CIERA is a collaboration of
University of Michigan
University of Virginia
Michigan State University
with
University of Minnesota
University of Southern California

University of Michigan School of Education
610 E University Av, Rm 1600 SEB
Ann Arbor, MI 48109-1259
734.647.6940 voice
734.763.1229 fax
ciera@umich.edu

www.ciera.org